Re-using an Argument Corpus to Aid in the Curation of Social Media Collections

Re-using an Argument Corpus to Aid in the Curation of Social Media Collections

Clare Llewellyn, Claire Grover, Jon Oberlander, Ewan Klein

School of Informatics
University of Edinburgh, Edinburgh, United Kingdom
c.a.llewellyn@sms.ed.ac.uk, grover@inf.ed.ac.uk, jon@inf.ed.ac.uk, ewan@inf.ed.ac.uk

Abstract

This work investigates how automated methods can be used to classify social media text into argumentation types. In particular it is shown how supervised machine learning was used to annotate a Twitter dataset (London Riots) with argumentation classes. An investigation of issues arising from a natural inconsistency within social media data found that machine learning algorithms tend to overfit to the data because Twitter contains a lot of repetition in the form of retweets. It is also noted that when learning argumentation classes we must be aware that the classes will most likely be of very different sizes and this must be kept in mind when analysing the results. Encouraging results were found in adapting a model from one domain of Twitter data (London Riots) to another (OR2012). When adapting a model to another dataset the most useful feature was punctuation. It is probable that the nature of punctuation in Twitter language, the very specific use in links, indicates argumentation class.

Keywords: Argumentation, Twitter Data, Curation

1. Introduction

There is a very large volume of data available from within the social media domain. To re-use this data it needs to be sensibly filtered, curated and archived (Kavanaugh et al., 2012). Organising this data by hand would be extremely time consuming, but automatic curation could mean that social media data could be gathered, organised, analysed and re-used on a large scale. As social media discussions can often take the form of a conversation or a debate it is proposed that extracting an argumentation structure from the data and using this to annotate it would be useful and assist in curation of this data.

The theory of argumentation formalises how humans disagree, debate and form consensus. It describes a structure for classifying discussions. This type of structure and the reasoning it supports is used widely in the fields of logic, AI and text processing (Mochales and Ieven, 2009). Although it is complex and not uniformly agreed upon, the general consensus is that an argument is composed of a claim, which is a statement of the position that the claimant is arguing for, and that this claim can be challenged with a counter claim. The claim / counter claim are backed up with premises, evidence that supports the claim (Toulmin, 2003; Walton et al., 2008).

Although implicitly understood, the structure of arguments can be difficult for humans to identify and describe consistently and this makes it difficult to perform argument mining (the automatic identification of these argument structures within text) (Mochales and Moens, 2011). Generally, this task uses an annotated argumentation corpus to train models to identify the linguistic features of argumentation and extract relationships between these features (Reed et al., 2008). There are few large scale argumentation corpora (Mochales and Ieven, 2009) and they tend to be composed of traditional media such as legal documents, parliamentary records, newspapers and journal articles (Hachey and Grover, 2005; Reed and Rowe, 2004; Teufel et al., 1999). High quality social media data sets annotated with argumentation structure are rare and should therefore be reused as widely as possible. Previous work within the social media domain includes using argument theory to monitor and assist in collaborative work learning (Rose et al., 2008) and using tweets classified in an argument structure to explore the rumors around the 2011 London Riots (Procter et al., 2011)

In this paper the initial aim is to show that automated methods can be used to classify social media text into argumentation types. This work follows the approach proposed by Rose et al. (2008) and compares results with that experimentation. In this work, a social media data set annotated with argumentation classes is used as training for the machine learning of these classes. Experimentation is conducted that compares the different machine learning algorithms for classifying further social media text and the accuracy of this approach is evaluated with respect to different features. An investigation of issues that arise due to natural inconsistency within social media data is conducted and the general pitfalls that may occur with this kind of data are identified. The classification model derived through machine learning for the initial dataset is then applied to data from another collection to determine if the model can be used more widely than the initial problem area.

2. Methods

2.1. Data

The data set that is used was taken from Twitter. Text that is posted in Twitter is different from many other content types as each 'document' is very short, it does not always use standard spelling or grammar and contains many specific

conventions associated with this specific medium such as retweeting, hashtags and emoticons. The data is noisy and covers a large number of parallel conversations all occurring at the same time. This type of data is attractive as it is freely available, in large volume, and is presented in clearly structured short conversational segments.

The initial data set used here is composed of hand-annotated tweets that describe the events that occurred during the London Riots in 2011. This data was originally gathered to investigate how social media is used in crisis situations (Procter et al., 2013). The data has also been used by Procter et al. (2011) to create a visualisation of how rumours spread through Twitter during the London Riots for the Guardian newspaper. Procter et al. (2013) developed a code frame which they used to describe the data that they gathered. A subsection of the data (7,729 tweets) is used here, the section described by Procter et al. (2013) using the 'rumours' code frame. In accordance with the code frame each tweet was post-annotated with a single code which expressed the type of argument within a simplistic argumentation structure (Table 1). This means that each tweet was a member of a single class within this structure. The codes were assigned by two human annotators with a third annotator arbitrating when there was disagreement. The intercoder agreement is stated to range between 89-96% (Procter et al., 2013).

Type Code	Class	Number of Tweets
1	Claim Without Evidence	2117
2	Claim With Evidence	3644
3	Counter Claim Without Evidence	689
4	Counter Claim With Evidence	268
5	Implicit Request for Verification	579
6	Explicit Request for Verification	0
7	Comment	384
8	Other	13
	Uncoded	35

Table 1: Number of tweets annotated with argument classes

The tweet data is in JSON format where the content is contained in attribute-value tuples containing a field describing the data type and a field containing the specific data. An example of the information available for each tweet is presented in Table 2. The body tuple contains the tweet text and the type tuple contains the argument class. In this case the type tuple is 2 which corresponds to **Claim With Evidence**.

Models derived from the riot data were applied to a secondary data set collected by the organiser of the Open Repositories 2012 conference (OR2012). The data (over 5,000 tweets) was collected in order to gauge the social media amplification of the event (Llewellyn et al., 2013). A small sub-section of this data (100 tweets) was annotated using the coding scheme in Table 1.

Name	Content
author	*****
body	RT @******: #londonriots oh my god - reports of tigersroaming around Primrose Hill #londonzoobreakin http://t.co/j2DjbOZ
id	683295
influence	72
parent	628335
time	1312836361
type	2

Table 2: Example Tweet

2.2. Machine Learning

A supervised machine learning model was trained and used to classify the tweets into the argument structure. Previously Rose et al. (2008) investigated using supervised machine learning to classify text from online discussion forums in order to improve collaborative learning within that forum— their experimentation was used as a template for this work. The Rose et al. (2008) experimentation compared different machine learning algorithms for classifying text into an argumentation classication system from Weinberger and Fischer (2006). The categories in the Riot Twitter corpus (Procter et al., 2013) do not directly match those used by Rose et al. (2008) and we therefore created a mapping, as shown in Table 3. As can be seen some of the codes from the Twitter scheme are combined to map to a single Weinberger and Fischer class and there are codes in the Weinberger and Fischer scheme that are not present in the Twitter corpus.

Both this and the Rose et al. (2008) work investigate which features are most useful in predicting the different argumentation classes using the TagHelper Tool, a corpus analysis environment built on top of the Weka machine learning toolkit (Rose et al., 2008; Witten and Frank, 2005). This tool provides various supervised machine learning algorithms. It allows users to extract features from text which are then used to create vectors which express that text. The features include **unigrams** and **bigrams** (with or without a stop list and stemming), **punctuation**, **line length**, and **part of speech bigrams**. It also allows the creation of rules that group and/or map certain features; in this case it is used to specify the most frequently occurring **non-topic specific** words.

Initially two questions were explored: which supervised machine learning algorithms were most suitable for this task and which features could most successfully distinguish the classes. The algorithms that were tested were Naive Bayes, Support Vector Machines and Decision Trees. Baselines were created using these algorithms with a limited set of top unigram features. Rose et al. (2008) used the top 100 for the discussion forum data and we used the top 82 for the Twitter data (as there was no significant improvement in the scores after this number). For each

Dimensions from Weinberger and Fischer	Coding Scheme used in the Twitter Corpus
Epistemic Activity	Comment Other
Micro-level of argumentation Macro-level of argumentation	Claim without evidence Claim with evidence Counter claim without evidence Counter claim with evidence
Social modes of co-construction	Implicit request for verification Explicit request for verification
Reaction to previous contribution	None
Reaction to script (prompt)	None
Quoted text	None

Table 3: Alignment of coding schemes

algorithm a score is reported for each class predicted.

To investigate the most useful features we have followed the steps identified by Rose et al. (2008) where unigram features are analysed in conjunction with several different additional feature sets; these sets include:

- Unigrams
- Unigrams and line length
- Unigrams and part of speech bigrams
- Unigrams and bigrams
- Unigrams and punctuation
- Unigrams and stemming
- Unigrams (rare words removed)

The results presented for this work take an average score for the accuracy across all the classes discussed above and are presented as a single value for each set of features.

The third stage of the work is an exploration into whether the model trained using the Twitter riot data set can be re-used to classify the arguments in a Twitter data set on a different subject. The data in this set is the tweets collected about the OR2012 conference.

It is expected that the unigram models trained on the riot data will not work well with data discussing a different subject. Therefore this experiment compares the performance of features that are dependent on text in the riot data set; unigram and bigram features with the performance of other features set independently (not in addition to unigrams as with the feature set analysis described above). The feature sets tested are:

- Unigrams and Bigrams
- Punctuation
- Punctuation and Line Length
- Punctuation, Line Length and Part Of Speech Bigrams
- Punctuation, Line Length and Non-Topic Words

One of the feature sets in the list above is **Non-Topic Words**; this feature set was made up of 100 words selected from a set of frequently used words extracted from a generic stop list (the type which is used to remove frequently occurring words when analysing text). This was used as a proxy for frequently occurring, non-topic specific words. The aim of this was to replace the use of unigrams, which were expected to perform badly when applying the riot model to the OR2012 set. Again the results presented are an average across all classes.

For each of the experiments the classication was evaluated by determining the level of agreement between the two classications of the data, automatic and manual. For each run of the data 10 fold cross-validation is conducted, data is randomly distributed into 10 sets and the results given are those averaged over the experimentation on the 10 different sets.

2.3. Metrics

Rose et al. (2008) use Cohen's kappa (κ) as a measure of agreement between the classes annotated by the human and those derived automatically. This measure represents the agreement between the annotators (in this case the human and the computer) modified to take account of the likelihood of the agreement occurring by chance. The likelihood of something occurring by chance would give a κ of 0.0. There is much disagreement about what represents a good κ score. Landis and Koch (1977) provide a commonly cited description: a score of 0 to 0.20 is slight, 0.21 to 0.40 is fair, 0.41 to 0.60 is moderate, 0.61 to 0.80 is substantial agreement.

In order to provide a comparison with the Rose et al. (2008) work the κ results are presented, but it is acknowledged that this is not a definitive description and there has been much discussion about whether this metric should be used at all (Powers, 2012). Therefore another metric, the *Matthew's Correlation Coefficient* (MCC) is provided. This metric is used because within both datasets some classes are represented more strongly than others. The MCC, to some extent, provides results that mitigate this issue. It is a measure used in machine learning to indicate agreement and is thought to work well even when the class sizes are very different. Again, a score of 0 would represent random agreement and 1 would be perfect agreement (Baldi et al., 2000).

Discussion Forum				Twitter (Incl. Retweets)			
Class	NB (κ)	SVM (κ)	DT (κ)	Class	NB (κ)	SVM (κ)	DT (κ)
Micro-level of argumentation	0.47	0.60	0.55	Claim	0.35	0.78	0.84
				Claim with evidence	0.68	0.83	0.86
Macro-level of argumentation	0.51	0.70	0.68	Counter claim	0.58	0.72	0.79
				Counter claim with evidence	0.42	0.47	0.84
Social modes of co-construction	0.38	0.48	0.49	Implicit request for verification	0.42	0.26	0.47
Epistemic Activity	0.42	0.53	0.47	Comment	0.44	0.30	0.49
				Other	0.63	0.45	0.36

Table 4: Classification of Online Discussion and Tweet Data - Unigram Performance
(Discussion forum data taken from Rose et al. (2008) some classes are not presented here as they have no equivalent)

3. Results and Discussion

3.1. Comparison of Machine Learning Algorithms

We found that in the initial data set (the riot data), classes of argument structure could be learnt and predicted to a high level of accuracy, as can be seen in Table 4. The machine learning algorithms gave better results for the argumentation classes (**Claim, claim with evidence, counter claim, counter claim with evidence**) than the **Implicit request for verification**, **Comment** or the **Other** classes. This is true for both this and the Rose et al. (2008) work. This suggests that there is a clear set of features that can be used to predict the argument classes, but there may be more variability in the **Comment**, **Implicit request for verification**, and **Other** data, making it more difficult to extract a good set of predictive features for these classes.

The algorithms performed with very different degrees of accuracy depending on the class predicted; for example, Naive Bayes was most able to predict the **Other** class, but least able to predict the argumentation classes. In general, we found that a Decision Tree gave the most consistent performance. Our results differed from Rose et al. (2008), who found that SVM was the most consistent performing algorithm. This suggests that the performance of the algorithms is dependent on the specific data set and that it can therefore be difficult to predict which algorithm will perform well. In this work, because Twitter data is of a similar nature to discussion forum data, we would have expected the most accurate algorithm to mirror the results of Rose et al. (2008), but it did not.

3.2. Feature Selection

The results presented in Table 5 are provided using the SVM algorithm (to mirror the Rose et al. (2008) work).Rose et al. (2008) found that when analysing the discussion forum data the additional features that improved performance, beyond unigram performance (0.48κ), were punctuation (0.52κ), part of speech bigrams (0.49κ) and stemming the unigrams (0.49κ). We found that with analysis of Twitter data in addition to unigrams (0.68κ) the most successful performance was the part of speech bigrams (0.85κ) feature set. Punctuation (0.79κ), bigrams (0.70) and line length (0.69κ) features also improved performance, but rare words gave no difference and stemming the unigrams actually reduced the performance (0.66κ).

These results suggest that the performance of additional features sets is, to some extent, dependent on the data sets used. But the comparison of results across the two data sets indicate that punctuation and part of speech bigram feature sets improve across both and would therefore be considered the most consistent. These feature sets will therefore be the focus of particular attention when adapting the model trained on the riot data to a different Twitter data set.

Features	Discussion Forum (κ)	Twitter Riot Data (κ)
Unigrams	0.48	0.68
Unigrams and line length	0.48	0.69
Unigrams and POS bigrams	0.49	0.85
Unigrams and bigrams	0.44	0.70
Unigrams and punctuation	0.52	0.79
Unigrams and Stemming	0.49	0.66
Unigrams and rare words	0.48	0.68

Table 5: Feature Selection -κ scores presented as an average of all classes as provided by TagHelper Toolkit (empty class explicit request for verification included in calculation)

3.3. OR2012 Data

Text dependent features (unigrams and bigrams) work well in training models to automatically classify the London Riots Twitter data (0.84 MCC) as seen in Table 6. When these models are applied to the secondary data set, the OR2012 Twitter data, human and automatic classification agree to a fair degree (0.25 MCC). However, when the resulting data is inspected, it is clear that this does not provide results that could be used without substantial human intervention.

As the results for text dependent features are not useful, the other features are investigated independently, as opposed to in addition to unigrams as with the previous work. In the previous feature selection work, the punctuation set performed most consistently across data sets. Therefore

	Riot Data			OR2012 Data		
Features	Correctly Classified (%)	κ	MCC	Correctly Classified (%)	κ	MCC
Unigrams, Bigrams	85.97	0.8123	0.8363	38.00	0.1709	0.2530
Punctuation	61.63	0.4681	0.5523	57.00	0.3869	0.4840
Punctuation, Line Length	61.68	0.4691	0.5530	58.00	0.3981	0.4840
POS Bigrams, Punctuation, Line Length	81.17	0.7479	0.7804	40.00	0.2141	0.2769
Non-Topic Words, Punctuation, Line Length	69.87	0.5892	0.6486	49.00	0.2980	0.3845

Table 6: Classification of Tweet data - Feature Selection - Riot Data and Open repository 2012 Data

punctuation is used as the base set of features with all other sets added to this.

It was found that no feature set improved upon the base punctuation set which gave a moderate agreement (0.48 MCC) although this was substantial improvement on the text dependent features (0.25 MCC). We believe that this is because the language in data chosen for the riot set is skewed towards particular topics. The initial corpus was collected to represent events that were rumored to have taken place during the riots. Some topics are more heavily represented within certain classes, therefore, words that indicate these topics are being used in the machine learning to identify the class that is over represented.

The results gained using the punctuation set gave a moderate agreement between human and machine (0.48 MCC). Again this does not provide results that could be used without human intervention. This feature set does perform better for some classes than others, and this is explored in the example section below. A reliance on a machine to automatically curate these tweets with an argument code in these classes may be possible, allowing a human to focus on the more complex cases.

It was surprising that the part of speech bigram features (0.28 MCC) and the non-topic words (0.39 MCC) did not improve performance but substantially reduced it (compared to the punctuation set (0.48 MCC)) as these features were chosen to try and avoid dependence on specific data sets. We hoped that these features would replace the use of unigrams when the model was adapted to an alternative data set. To investigate why this did not work, some specific examples are considered below.

3.4. Issues with the Twitter Data

In our experimentation several observations were made about working with Twitter data and some allowances had to be made for this specific type of data. The next sections discuss these problems.

3.4.1. Retweets

One observation, and a specific issue in Twitter, is that people retweet previous tweets. The London Riot data set is a collection of tweets which show how the rumors about the London riots proliferated through Twitter. The sampling of this data is focused on collecting retweets to show the proliferation and this means that they are a large number in this data set. The OR2012 data collects all tweets which contain the #OR2012 so does not discriminate for or against retweets. Several previous studies (Castellanos et al., 2011; **?**) describe how results can be skewed by retweets. As it is a particularly significant aspect of this data, a reflection of the purpose of collecting the London Riot data and a different collection strategy to the OR2012 data, it is discussed in more detail here.

Within 7,729 original tweets there are only 2,786 individual unique tweets. If this repetition remains in the data, the features identified in the machine learning would closely reflect the language in the highly repeated tweets, resulting in over-fitting to the riot data. To ensure that the model is useful beyond this specific data set it was pre-processed to remove the repetition.

In Table 7 this over-fitting can be seen, the 'without retweets' set gives lower κ values than the 'including the retweets' for the riot data for both the unigrams punctuation and line length feature set and the punctuation and line length feature set. When the model is re-used for the OR2012 data these result show a higher agreement for the model that is constructed 'without retweets'. These results indicate that it is important, if a set is to be reused, to remove the repetition caused by retweets.

3.4.2. Class size

Within the smaller classes in the London Riot data, some topics are more heavily represented than others; therefore words that indicate these topics are being used in the machine learning to identify the class. For example, the word *hospital* was a strong indication of the class **implicit request for verification** as people were tweeting to ask what was happening at a children's hospital. An example of a tweet that does this is:

```
They can't really be trying to get into
the childrens hospital can they?
#birminghamriots
```

This is probably one of the reasons that the machine learning model does not transfer well to a second data set when unigram and bigram features are used.

Training Data Set	Features	Riot Data (κ)	OR2012 Data (κ)
With Retweets	Punctuation and Line Length	0.59	0.33
Without Retweets	Punctuation and Line Length	0.47	0.401
With Retweets	Unigrams, Punctuation and Line Length	0.92	0.22
Without Retweets	Unigrams, Punctuation and Line Length	0.79	0.25

Table 7: Classification of Tweet data - Using and not using Retweets

Within both datasets, some classes are represented more strongly than others; for example there are many more **claims** and **claims without evidence** than the other classes (see Table 1). When providing an average across all classes, a method that gives a good result in the larger classes may overshadow a good result in a smaller class. When we evaluate the results it is useful to use a measure that takes account of different class sizes. The *Matthew's Correlation Coefficient* (MCC) was thus used to provide results that, to some extent, mitigate this issue.

It can be seen that the scores in Table 6 provide a comparison of κ and MCC scores. In particular, it is interesting to note that the MCC measurement accentuates the increase in agreement when using the punctuation (0.48 MCC) feature set more than the other sets. This indicates that this feature set may work better for the smaller classes than the other feature sets.

3.4.3. Examples

The results for the OR2012 Twitter data using punctuation show some promise for being able to identify argumentation classes using a punctuation feature set (0.48 MCC): this would be judged as moderate agreement between human and automatic classification. Where the model does well is considered in more detail here via some examples. The model built using the punctuation features is good at predicting whether a claim does or does not have evidence, for example a correctly classified **Claim with Evidence**:

```
RT @moragm23 : Yale University Arabic
and Middle Eastern Electronic Library
- interesting page turner plugin #or2012
http://t.co/QVRHd2GZ
```

In both data sets, the evidence is generally a link to a website or a digital photograph. These links therefore include indicative punctuation marks, in particular the forward slash, which is rarely used outside a link.

The difference between **claim** and **counter claims** is more difficult to determine. Here is a tweet from the class **claim** that was correctly identified using the punctuation feature set (whereas other features sets predict the class incorrectly):

```
Different roles, different repositories:
backup, sharing, archiving. #or2012
```

This is also classified correctly using the feature set of punctuation, line length and part of speech, but it is classed as a **claim with evidence** when using unigrams and bigrams. A **counter claim** is identified correctly using the punctuation feature set:

```
@NamesProject esearchers are
pussy-cats:dont like sticks or
carrots. maybe institution respond
to both more effectively.#or2012
```

But this is classified as a **claim** using the feature set of punctuation, line length and part of speech, and as a **claim with evidence** when using unigrams and bigrams.

In general, it is very difficult to determine why the machine learner makes the class prediction in each of these cases, as it involves a complex mixture of features. However, it is possible to surmise why it made a mistake in certain instances when there are particularly strong features. For example, this tweet was classed by the human as a **claim**, but using the punctuation, line length and part of speech feature set as an **implicit request for verification**.

```
RT @lljohnston: What are the right
licenses for deposit of software for
preservation? OSI approved licenses.
Recording of license a must for ingest.
#or2012
```

In this case it is most likely that this has been misclassified because it contains a question mark and a WP VBP part of speech bigram (what are), which are both indicators of the **implicit request for verification** class.

In other cases it is almost impossible to tell why the mistake has been made, for example the tweet below is classed as a **counter claim** by the human but as a **claim** using the punctuation, line length and part of speech feature set.

```
@lljohnston You forgot procrastination
:) #or2012
```

The part of speech bigram PRP VBD (You forgot) is a strong predictor of the **counter claim class**, yet it is still classed incorrectly.

4. Conclusions

Supervised machine learning was performed on a Twitter data set to identify arguments within text. In this difficult task, we found that these methods can successfully distinguish between different types of argument. Encouraging results were found in adapting a model from one domain of

Twitter data (London Riots) to another (OR2012). Twitter data contains a lot of repetition (because of retweets), and we found that this repetition can cause the machine learning algorithms to overfit to this data. We also noted that when learning argumentation classes, we must be aware that the classes will most likely be of very different sizes and so we needed to account for this when we analysed the results.

We discovered that when adapting a given model to another dataset the most useful feature was punctuation. We hoped that using a small set of non-topic specific words would outperform the language independent features, but they did not. It is probable that the nature of punctuation in Twitter language (the very specific use in links) indicates argumentation class. The results gained using the punctuation set gave a moderate agreement, but do not provide results that could be used without human intervention. An inspection of the data has indicated that in certain cases it may be possible to rely on a machine to automatically curate those tweets with a subset of argument codes, allowing a human to focus on the more complex cases.

One of the drawbacks in using machine learning in this manner is that it is difficult to engineer which features should be used. If the feature engineering could be done more effectively, it may be possible to identify the more appropriate parts of speech bigrams and it may be feasible to improve on the current results.

Acknowledgements

We would like to thank Nicola Osborne for collecting and providing the use of the OR2012 Tweet corpus. We would also like to thank Rob Procter and Analysing Social Media Collaboration for access to the London Riots Twitter corpus.

5. References

Pierre Baldi, Sren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Malu Castellanos, Umeshwar Dayal, Meichun Hsu, Riddhiman Ghosh, Mohamed Dekhil, Yue Lu, Lei Zhang, and Mark Schreiman. 2011. Lci: a social channel analysis platform for live customer intelligence. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1049–1058. ACM.

Ben Hachey and Claire Grover. 2005. Sequence modelling for sentence classification in a legal summarisation system. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 292–296. ACM.

Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Donald J. Shoemaker, Apostol Natsev, and Lexing Xie. 2012. Social media use by government: From the routine to the critical. 29(4):480–491.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):pp. 159–174.

Clare Llewellyn, Nichola Osborne, Ewan Klein, and Miranda Taylor. 2013. An analysis of professional exchange and community dynamics on twitter around the #or2012 conference hashtag - a session at twitter and microblogging: Political, professional and personal practices.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the echr. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 21–30, New York, NY, USA. ACM.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

David M. W. Powers. 2012. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, page 345355. Association for Computational Linguistics.

Rob Procter, Farida Vis, and Alexander Voss. 2011. Riot rumours: how misinformation spread on twitter during a time of crisis. *the Guardian*, December.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. 16(3):197–214.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 14:961980.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*, pages 91–100.

Carolyn Rose, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, page 110117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press, updated edition edition, July.

Douglas N Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press, Cambridge; New York.

Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95, January.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, 2 edition, June.